# Social-consensus feedback as a strategy to overcome spontaneous gender stereotypes

Article (Accepted Version)

# Social-consensus feedback as a strategy to overcome spontaneous gender stereotypes

Eimear Finnegan, Alan Garnham and Jane Oakhill

School of Psychology, University of Sussex, Brighton, England

This research was supported by funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n° 237907. Correspondence concerning this article should be addressed to Alan Garnham, School of Psychology, University of Sussex, Brighton, BN1 9QH, England (e-mail: a.garnham@sussex.ac.uk).

Social-consensus feedback as a strategy to overcome spontaneous gender stereotypes


## Abstract

Across two experiments the present research examined the use of social-consensus feedback as a strategy for overcoming spontaneous gender stereotyping when certain social role nouns and professional terms are read. Participants were presented with word pairs comprising a role noun (e.g. surgeon) and a kinship term (e.g. mother), and asked to decide whether both terms could refer to the same person. In the absence of training, participants responded more slowly and less accurately to stereotype incongruent pairings (e.g. surgeon/mother) than stereotype congruent pairings (e.g. surgeon/father). When participants were provided with (fictitious) social consensus feedback, constructed so as to suggest that past participants did not succumb to stereotypes, performance to incongruent pairings improved significantly (Experiment 1). The mechanism(s) through which the social feedback operated were then investigated (Experiment 2), with results suggesting that success was owing to social compliance processes. Implications of findings for the field of discourse processing are discussed.

Considerable evidence suggests that readers often make gender inferences in text comprehension when explicit gender information is lacking (e.g. Carreiras, Garnham, Oakhill, & Cain, 1996; Duffy & Keir, 2004; Garnham, Oakhill, & Reynolds, 2002; Irmen, 2007; Kreiner, Sturt, & Garrod, 2008). Such inferences typically follow the use of social or occupational role nouns that have a strong gender bias, but that are not grammatically marked for gender e.g. the term *beautician* is strongly female-biased while the term *builder* is strongly male-biased. While grammatical gender languages can largely avoid gender stereotypic inferences by employing gender specific personal nouns to convey maleness and femaleness (e.g. *le musician* 'the [male] musician'/*la musicienne* 'the [female] musician' in French versus *the musician* in English), this is rarely possible in English. Instead, inferences based on stereotypical biases play an important role in building a cognitive representation of gender and, once established, are very difficult to overcome. This can result in processing difficulties when gender-related expectancies clash with explicitly stated gender information.

Reynolds, Garnham and Oakhill (2006) gave participants a slightly adapted version of the now well known 'surgeon riddle' referenced by Sanford (1985). In this riddle, a father and son are involved in a car accident where the father dies but the son is taken to hospital for an operation. However, once there, the surgeon looks at the boy and exclaims "Oh my god, that is my son!" (Sanford, 1985: "I can't do this operation. This boy is my son."). When readers are asked how this can be, they typically infer that the surgeon is male and, despite knowing that the boy's father is dead, fail to override this inference so as to successfully conclude that the surgeon is the boy's mother. Indeed, Reynolds and colleagues report that 75% of readers who had not previously seen the text failed to resolve the inconsistency and update the gender of the surgeon in their mental representation. However, participants who received a different version of this riddle with the term surgeon replaced by nurse showed no difficulty in solving it (Experiment 1A). Moreover, when these findings were followed up using a self-paced reading task (Experiment 2), reading times for the final clause (which contained the gender-biased role noun) were found to be 1,000ms slower on

the original surgeon riddle than the stereotype consistent version. This processing delay is much longer than is typically found in previous experiments that require minor accommodations to mental representations, i.e. approximately 200ms (e.g. Carreiras et al., 1996, Experiment 1; Haviland & Clark, 1974).

The above evidence led Reynolds et al. to conclude that inferring gender from stereotyped role-names is at least in some part an automatic process, with the experiments highlighting how entrenched such gender inferences are and the substantial processing difficulties that they induce. Such incorrect inferences based on stereotypicality biases clearly have negative implications for discourse comprehension but are also a pervasive example of how language contributes to the maintenance and propagation of gender stereotypes in English, by artificially constraining the roles on offer to men and women. Therefore, the current article investigates how spontaneous gender stereotypes can be overcome through the use of (fictitious) social consensus feedback. This feedback involves presenting participants with social norm information relating to role-based gender stereotypes in an attempt to sway their attitudes towards the perceived (gender-fair) attitude of their peers.

A sizeable body of research has now been devoted to the influence of other people's beliefs on an individual's own beliefs, with evidence emerging that perceivers frequently modify their intergroup attitudes and behaviours in order to align with those modelled by members of groups that they value. For example, informing participants that stereotyping is not typical of their in-group has previously been found to reduce stereotyping against groups such as racial minorities (e.g. Stangor, Sechrist, & Jost, 2001; Wittenbrink & Henly, 1996) and people suffering from obesity (Puhl, Schwartz, & Brownell, 2005) while, conversely, research has documented that discrimination against racial minorities and women is more tolerated when a racist or sexist joke has just been heard (Ford & Ferguson, 2004; LaFrance & Woodzicka, 1998).

Various theories have been offered to explain this peer influence on intergroup prejudice, now thought to be driven by the basic human goals of understanding, social connection, or

alternatively, self-definition (Paluck, 2011). For instance, Social Reality Theory posits that striving for understanding and connection pushes people to validate their experiences with others and to display actions and beliefs that others value (Hardin & Conley, 2000). Similarly, Group Norms Theory suggests that people assume the perceived attitudes and behaviours of others who exemplify admirable in-group identities in an effort to socially connect with the group (Crandall, Eshleman, & O'Brien, 2002; Kelman, 1958; Sherif & Sherif, 1953). While both of these theories predict that individuals would adjust their behaviour and attitudes in line with those of the valued reference group, other theories predict a contrasting pattern of results. Specifically, Deviance Regulation Theory (Blanton & Christie, 2003) claims that people may reject the perceived attitudes and behaviours of their peers as a means of self-definition, while the Focus Theory of normative conduct predicts that peer values will only influence individuals when they are made salient (Kallgren, Reno, & Cialdini, 2000). Overall, given past success in using social norm information as a strategy for overcoming prejudice in other domains, we posit that participants are more likely to be influenced by the perceived attitudes of their peers than to reject them as a means of self-definition.

In a similar vein, Prentice and Miller (1993) posit that individuals will experience discomfort if they perceive their attitudes to be different from the normative attitude of their peer group. However, this discrepancy can be resolved in three ways (1) by moving an individual's personal attitudes towards that of the perceived norm, (2) bringing the norm closer to the individual's attitude, or (3) complete rejection of the group. Prentice and Miller maintain that the most straightforward way for an individual to reduce a perceived discrepancy in attitude is to bring their private attitudes in line with those of the group norm. Furthermore, as regards stereotyping, Stangor et al. (2001) propose three reasons why individuals should be particularly likely to be swayed by the opinions of others on this issue (1) the accuracy of stereotypes is difficult to assess objectively (2) stereotyping is a socially sensitive topic and (3) people are likely to be highly motivated to learn about the traits of individuals from different social groups.

However, while many studies suggest people reform their attitudes and beliefs so as to mirror those of their peer group, less research has explored how well people can initially identify these social norms. In fact, it is now clear that they can make significant errors in their estimation of opinions held by others (e.g. Prentice & Miller, 1993). The use of fictitious norm information as a stereotype reduction strategy in this, and other, research therefore benefits from the fact that people are often poor at estimating social norms, yet are strongly influenced by what they perceive these norms to be.

With the above information in mind, Experiment 1 was designed to investigate the effect of fictitious social consensus feedback on levels of gender stereotype endorsement on a straightforward judgement task. This task was originally devised by Oakhill, Garnham and Reynolds (2005) who conducted a series of six experiments investigating gender biases associated with single role nouns and the extent to which such bias information could be overcome. Participants were required to judge quickly whether two terms presented on screen could refer to one person. The terms comprised a role noun (that was either definitionally gendered or stereotype biased e.g. *princess, beautician* respectively) and a kinship term (that was definitionally gendered in all cases e.g. *brother, sister*). In order to successfully respond, participants were required take definitional gender into account (i.e. a brother is always male) but to dismiss stereotypical gender (e.g. that most beauticians are female).

Presentation of stimuli and the details of the instructions were varied across the studies, but in all studies participants consistently rejected gender incongruent word pairings (e.g. *beautician/brother*) more frequently than gender congruent pairings (e.g. *beautician/sister*). This pattern was still evident (although to a lesser extent) when participants were explicitly provided with a strategy to aid performance with incongruent pairings; they were reminded that nowadays many jobs can are not clearly marked for gender (i.e. have ‘man’, ‘woman’ or ‘ess’ in the title), and that they should carefully consider whether the presented role could be occupied by a man, woman or both (Experiment 4). Oakhill et al. concluded that there is likely an automatic component to

responding, as participants still struggled to suppress the gender stereotype information associated with the role nouns, despite it being counter-productive to task performance.

Other strategies have also been proposed to overcome stereotype biases to social/occupational role nouns. These terms are frequently used in sentence comprehension studies using a match/mismatch paradigm in which a stereotyped term is followed by gender congruent or incongruent information. In the former condition processing is typically unproblematic (e.g. the *builder* went to work although *he* was not feeling well), whereas difficulty arises in the latter version (e.g. the *builder* went to work although *she* was not feeling well). This difficulty in integrating the unexpected gender information into the reader's representation of the text is often conveyed through slower judgement or reading times relative to the gender matching condition (e.g. Carreiras et al., 1996; Duffy & Keir, 2004; Garnham, Gabriel, Sarrasin, Gygax, & Oakhill, 2012; Garnham et al., 2002; Irmen, 2007; Kennison & Trofe, 2003; Kreiner et al., 2008). However, while these gender stereotype biases can be successfully overcome by establishing the sex of a character *before* a role noun is encountered (e.g. Duffy & Keir, 2004; Kreiner, et al., 2008; Lassonde & O'Brien, 2013), such an approach is clearly not always practicable.

But what about strategies to overcome stereotyping more broadly, outside of the domain of language processing? Social psychologists have typically sought to reduce stereotypes by changing attitudes and behaviour within people's volitional control using a combination of awareness and effort (Dasgupta & Greenwald, 2001). For instance, this has been achieved by replacing automatic, culturally stereotypic responses with more considered responses that reflect personal beliefs (Devine, 1989; Monteith, 1993; Monteith, Devine, & Zuwerink, 1993) or by encouraging the suppression of negative stereotypes (Macrae, Bodenhausen, Milne, & Jetten, 1994). However, more recently cognitive psychologists have sought to address stereotype reduction by targeting automatic stereotypes using strategies such as stereotype negation/counter-stereotype affirmation training (e.g. Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008; Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000), mental imagery (Blair, Ma & Lenton, 2001), or through use of a counter-stereotype

expectancy strategy (Blair & Banaji, 1996).

Indeed, methods for overcoming both automatic and controlled stereotypes are of relevance to the current work as the judgement task of Oakhill et al. (2005) involves both. While stereotype activation takes place automatically and stems from increased cognitive accessibility of traits/features connected with a specific group, stereotype application is typically under the conscious control of a perceiver and involves the actual use of stereotypes in response to a group member (Kawakami, Dovidio, & van Kamp, 2005; see also Schneider & Shiffrin (1977) and Shiffrin & Schneider (1977) for a more detailed discussion of automatic and controlled processing). In the present context, participants must first overcome stereotype activation upon being presented with a gender-biased role noun to ultimately make the correct response. Therefore, in using this judgement paradigm of Oakhill and colleagues, the current article is ultimately concerned with achieving a reduction in stereotype *application*. If participants can learn to overcome spontaneous biases in a laboratory-based training, such interventions could have important implications for text processing and discourse comprehension more broadly.

**Overview of studies**

In two studies we investigated the influence of social consensus information on the responses of participants to gender-biased role nouns. Using the judgement task of Oakhill et al. (2005) participants were presented with three blocks of stereotype judgement trials, with feedback based on social norm information provided in Block 2 only. In Experiment 1 this feedback ostensibly indicated the percentage of students in a previous study that agreed with the participant's judgement. However, although presented to participants as true and accurate feedback, in reality, it was fictitious and manipulated so as to suggest that gender stereotype endorsement was very infrequent among the participant's peer group of fellow university students. It was hypothesised that participants would modify their responses towards the perceived attitudes of their peer group and display lower levels of stereotype application in Block 2. It was further hypothesised that this improved performance would be maintained in Block 3 (despite the removal of the feedback), with

participants investing continued effort to adapt their responding to the social norms they were presented with in Block 2.

Experiment 2 sought to identify the mechanisms through which the social feedback induced lower levels of stereotyping in Experiment 1: i.e. whether it worked through social compliance mechanisms or by alerting participants to the issue of stereotype bias through the use of majority feedback (and essentially reminding them that nowadays men and women can occupy many of the same roles, despite stereotype biases). In this way, Experiment 2 was largely similar to Experiment 1 but with the feedback adapted so as to suggest that stereotypic responding was *endorsed* by previous participants. As a result, it could be investigated whether participants (a) 'complied' with the feedback provided and responded in line with their peer group (thus failing to reduce stereotyping across blocks) or (b) were alerted to the issue of stereotype biases through the feedback provided, and consequently deemed counter-stereotypes pairings acceptable (thus resulting in a reduction of stereotyping across blocks).

## Experiment 1

### Method

**Participants.** Thirty-six students (17 male, 19 female) from the University of Sussex took part in this experiment. Participants' ages ranged from 18 to 30 years (*M*: 19.61; *SD*: 2.81). They received either £6 or 4 course credits for taking part in the session which lasted approximately 45 minutes.

**Materials & Design.**

***Gender-biased role nouns.*** Gender-biased role nouns were chosen from norms compiled by Gabriel, Gygax, Sarrasin, Garnham, and Oakhill (2008). The selected items were those rated as being most highly male-biased (e.g. bricklayer, president), most highly female-biased (e.g. beautician, fortune teller) or neutral (e.g. pedestrian, proof reader), with 12 exemplars chosen in each of these three conditions. A full list of the stereotyped terms used is provided in Appendix A.

The original ratings reveal that the bias scores of the 12 male-biased items extend across a narrower range (11.10% from strongest ($M$ = 88.24%) to weakest (77.14%) bias rating) than the bias ratings of the 12 female-biased items (17.55% % from strongest ($M$ = 13.27%) to weakest (29.22%)), $t$ (22) = 3.53, $p$ = .002. This suggests that, on the whole, the female-biased items were not judged as being as strongly stereotype-biased as the male items, with ratings of the former closer to neutral. For this reason, participants may show less difficulty in overcoming stereotype biases to female-biased role nouns relative to male-biased nouns[1]. Finally, ratings of the neutral terms extended across a very narrow range of 5.29% (around the 50% neutral mark; $M$ = 52.94 to 47.65%). Combined with the fact that participants should have ample experience of males and females fulfilling neutral roles, this narrow range of bias ratings is another reason that these terms should prove unproblematic for participants.

***Kinship terms.*** As in Oakhill et al. (2005), six kinship terms (three male, three female) were also selected to be used as one of the terms in the word pairs. These terms were *father, mother, brother, sister, uncle, aunt*. Importantly, these words incorporate a specific gender into their definitions e.g. the term *father* can only refer to males.

***Critical[2] word pairs.*** Word pairs were formed by combining the 12 male-biased, 12 female-biased and 12 neutral role nouns with the 6 kinship terms to produce a set of stereotype congruent, stereotype incongruent and neutral pairings. In the congruent condition, male and female stereotyped role names were paired with a kinship term of congruent definitional gender – for example, *pilot/brother* or *nurse/sister*. In the incongruent condition the stereotyped terms were paired with a kinship term of incongruent definitional gender – for example, *nurse/brother* or *pilot/sister*. Finally, in the neutral condition, neutrally rated role nouns were paired with each of the male and female kinship terms to create neutral word pairs – for example, *artist/father* and *artist/mother*. Overall, each of the 12 male-biased, female-biased and neutral role terms was teamed once with each of the six kinship terms resulting in 72 word pairs in each of the three congruence conditions, totalling 216 critical trials.

***Filler trials.*** Two-hundred and forty filler trials were also created, made by pairing the 6 kinship terms with role nouns that are also gender-specific by definition. In this way, filler trials were gender unambiguous pairings to which participants could respond *yes* or *no* to with relative ease and certainty. The selected role nouns were either explicitly marked for gender (e.g. waitress, or policeman), official titles (e.g. count or countess) or other terms that carry gender as part of their meaning (e.g. lady or husband). These role nouns were sourced from rating studies conducted by Hamilton (2008) and Kennison and Trofe (2003). A full list of the filler terms used is provided in Appendix B.

***Item overview.*** In total, participants were presented with 456 word pairs, divided into three equal blocks of 152 trials. Each of the stereotyped terms appeared twice in each block, once with a male kinship term and once with a female kinship term i.e. in both a congruent and an incongruent condition. The six kinship terms were counterbalanced so as to appear within the critical items an equal number of times in each block. Altogether 276 items, including all critical items, were intended to elicit a *yes* response while 180 required a *no* response.

***Social consensus feedback.*** Social consensus feedback was presented to participants as a strategy aimed at reducing levels of gender-stereotype application. This feedback was provided after each response in Block 2 of the judgement task, and consisted of a single sentence stating the percentage of University of Sussex students in a previous study run by the experimenter who agreed with the participant's judgement e.g. '*_% of previous students agreed with you'*.

As mentioned above, this feedback was in fact fictitious and constructed so as to suggest that the vast majority of previous participants accepted stereotype incongruent word pairs as warranting *yes* responses (i.e. as being perfectly acceptable). In this way, the social feedback sought to endorse gender-fair responding and highlight any discrepancy between a participant's response and the peer group norm. For example, if a participant responded that both terms of a stereotype incongruent word pair (e.g. *carpenter/sister*) could not refer to one person, feedback indicated that a

number between 2% and 5% of previous students agreed with this judgement. Conversely, if a participant judged such a pairing as acceptable, feedback indicated that a number between 95% and 98% of previous students agreed with the participant's choice, thereby reinforcing non-stereotypic responding. Note that the extreme, narrow range of feedback used was chosen so as to strongly and consistently convey that previous participants did not respond in a stereotyped manner.

A specific range of social consensus feedback was created for each of the three congruency conditions (see Appendix C), with exact figures within this specified range counterbalanced across pairings (e.g. with the stereotype incongruent trials, the figure 95% was presented an equal number of times as 96%, 97% and 98% in response to correct judgements). Aside from the stereotype incongruent trials, feedback to word pairs in all other congruency conditions (stereotype congruent, neutral, definitionally matching and definitionally mismatching word pairs) was loosely based on real data from other studies by the present authors (Finnegan, Oakhill & Garnham, 2014) with strong endorsement of correct responses and rejection of incorrect responses[3].

**Design.** In the judgement task, terms were presented one at a time in the centre of a computer screen using using E-Prime 2.0 software (Schneider, Eschman & Zuccolotto, 2002). A role term was first displayed for 1000ms, followed immediately by a kinship term (inter-stimulus interval of 0), which remained on-screen until a response was made. At this point (in Block 2 only), feedback immediately appeared on-screen (0 delay) and remained for 1,000ms. Finally, there was a 500ms delay before onset of the next trial. The word pairs were divided into three fixed sets to form the blocks of the experiment, while the sequence in which these blocks were presented to participants was counterbalanced. Within each block, trial order was randomised separately for each participant, using the standard E-Prime procedure. A Psychology Software Tools (E-prime manufactured) button box was used for responding, with one button clearly marked Y for *yes* and another N for *no*. Participants made a judgement about every word pair. The proportion of correct answers and response time of judgements to correct trials were analysed.

**Procedure.** Participants were tested individually in a quiet laboratory. Onscreen instructions informed them to read each pair of words and decide (without excessive deliberation) whether the two terms could apply to the same individual. These instructions provided the participants with two examples of such (definitional) word pairs – one that required a *yes* response and one that required a *no* response. Participants were further informed that they would receive feedback in the second block of judgement trials, and explained what this feedback entailed. The instructions and examples were then repeated verbally. Finally, in both conditions, a short practice session using a representative sample of fillers and critical word pairs was given to familiarise the participants with the experimental task. This consisted of eight trials and involved role terms that were not subsequently used in the experimental blocks.

After the experiment, a comprehensive debriefing session was held in which participants were informed that the feedback information was entirely fictitious. They were then reassured that, in reality, stereotype biases occur much more frequently than the feedback suggested and that there was no evidence that they were stereotyping to a greater extent than their peers.

## Results

**Data screening.** In this analysis, data for word pairs that contained the neutral term *adolescent* were excluded as accuracy of responses to such pairs was low, resulting in only 76% correct responses in Block 1 compared to > 90% accuracy for all other neutral role nouns. On reflection, this finding may be due to age considerations as opposed to gender stereotyping - the term *adolescent* typically refers to an individual in their teens and was paired with kinship terms that generally imply an older generation e.g. *uncle, aunt, mother, father*. This resulted in word pairs such as *adolescent/father*, which proved more difficult for participants to accept as correct than *adolescent/brother*, despite both being possible combinations. This resulted in the removal of 1.32% of the data.

**Analysis.** Across all experiments, accuracy of judgements and response times (RTs) were analysed using two mixed ANOVAs on the correct responses: firstly with participants treated as the random variable and secondly with items treated as the random variable. In the by-participants analysis ($F_1$), the mixed ANOVA had three repeated factors – stereotype bias of the role name (Stereotype: Male/Female/Neutral), gender of the kinship term (Kinship term gender: Male/Female) and block of trials (Block: Block1/Block2/Block3). Participant sex[4] (Male/Female) was included as between-subject factors. In the by-items analyses ($F_2$), Stereotype was included as a between-items factor while Kinship term gender, Block, and Participant Sex were included as within-item variables. In both sets of analyses, where sphericity was not satisfied, Greenhouse-Geisser (when $\varepsilon < 0.75$) or Huynh-Feldt ($\varepsilon > 0.75$) corrected degrees of freedom and *p* values are presented (as recommended by Girden, 1992). With all paired *t*-tests, within-subject or within-item effect sizes were estimated using Cohen's *dz* while with the independent-samples *t*-tests, estimates of between-subject or between-item effect sizes were estimated using Cohen's *d*. Finally, all graphs show the by-participant as opposed to by-item data in line with the more common practice in the literature.

***A note on Congruency.*** It is important to note that an interaction of Stereotype by Kinship term gender is equivalent to a main effect of Congruency as it is the combination of the levels of these two factors that give rise to the three critical conditions – congruent, incongruent and neutral. As such, from this point forward, all Stereotype by Kinship term gender interactions are referred to as effects of Congruency (though primarily in relation to the male and female stereotyped terms).

**Accuracy.** The main question of interest was whether performance to stereotype incongruent pairings improved across blocks when social consensus feedback was provided. A main effect of Congruency was revealed, $F_1$ (1.04, 35.24) = 16.30, $p < .001$; $F_2$ (2, 32) = 106.43, $p < .001$, driven by significantly lower accuracy to stereotype incongruent word pairs ($M$ = 79.9%), than to stereotype congruent ($M$ = 99.0%) and neutral pairs ($M$ = 96.4%). A significant interaction of Block by Congruency also emerged, $F_1$ (2.23, 75.96) = 3.08, $p = .046$; $F_2$ (4, 64) = 5.89, $p < .001$. As can be seen in Figure 1, this interaction was driven by a steady increase in accuracy of

stereotype incongruent pairings across blocks, totalling a 6.14% increase from Block 1 to Block 3. Due to ceiling effects, much smaller improvements in the accuracy of neutral and congruent conditions were found across blocks (1.9% and 0.6% respectively), both of which had very high accuracy from the outset.

**--insert Figure 1 about here--**

Paired samples *t*-tests[5] revealed that the aforementioned increase in accuracy to stereotype incongruent pairings across Blocks 1-3 was significant, $t_1$ (35) = 2.09, $p$ = .022, $dz$ = .35; $t_2$ (23) = 4.26, $p$ < .001, $dz$ = .87. It is posited that this accuracy improvement is due to the social feedback manipulation in Block 2 of the judgement task, thus providing support for the use of this feedback as a useful stereotype reduction strategy. However, by the end of the experiment, accuracy on stereotype incongruent word pairs remained significantly lower than that on stereotype congruent pairings, $t_1$ (35) = 3.69, $p$ = .001, $dz$ = .62; $t_2$ (23) = 9.80, $p$ < .001, $dz$ = 2.0, and neutral word pairs, $t_1$ (35) = 3.83, $p$ = .001, $dz$ = .64; $t_2$ (30.91) = 7.59, $p$ < .001, $d$ = 2.73. It can therefore be concluded that, despite a significant increase in accuracy to stereotype incongruent word pairs from Block 1 to Block 3, the social consensus feedback did not completely succeed in eliminating gender-biased responding.

**Response times.** Response times for all errors of judgement were excluded from the data set (representing 9.67% of the data) along with extreme response times, below 150ms and above 4,000ms (representing a further 2.77%), totalling a loss of 12.44% of the data. Next, the Participant by Block mean was calculated for each participant. Data points 2.5 standard deviations above or below the Participant by Block mean were replaced with the relevant upper or lower cut off point (a further 4.31% of the data).

A main effect of Congruency was found, $F_1$ (1.67, 56.82) = 18.31, $p$ < .001; $F_2$ (2, 32) = 12.47, $p$ < .001, with fastest RTs in response to stereotype congruent ($M$ = 824ms) and neutral word pairs ($M$ = 838ms), while RTs to incongruent pairings were considerably slower ($M$ = 967ms).

Contrary to expectations, no interaction of Congruency by Block was found, $F_1$ (2.82, 95.70) = 1.80, $p$ = .709; $F_2$ (3.78, 60.47) = 1.48, $p$ = .222. Indeed, Figure 2 illustrates that RTs in each of the three congruency conditions produced a relatively similar pattern of results. The sharpest fall in RTs was found with the stereotype incongruent pairings, decreasing a significant 263ms across Blocks 1-3, $t_1$ (35) = 3.95, $p < .001$, $dz$ = .66; $t_2$ (23) = 9.22, $p < .001$, $dz$ = 1.88 (versus 128ms for congruent pairings and 168ms for neutral pairings). While the social feedback initially had little effect on speed of responding to incongruent pairings in Block 2, it appeared to greatly aid subsequent performance in Block 3[6]. This delayed impact of the feedback on RTs is perhaps a result of the long sentence format of the consensus feedback, as increased processing time is likely to have been required before any effects of the training were evident.

**--insert Figure 2 about here--**

As RTs improved from Block 1 to Block 3 in all congruency conditions, the RT data provides provisional evidence for the use of social consensus feedback as a strategy for reducing the effects of gender stereotype activation. However, a significant difference between the RTs of stereotype congruent and incongruent pairings remained in Block 3, $t_1$ (35) = 2.07, $p$ = .046, $dz$ = .35; $t_2$ (23) = 3.0, $p$ = .006, $dz$ = .61. Again, this significant difference indicates that the social feedback training did not succeed in fully eradicating the stereotyping effect.

**Fillers - Accuracy.** Performance on filler trials was somewhat variable in Experiment 1. An average of 97.04% accuracy was found across conditions in response to the definitionally matching word pairs (e.g. *host/father*), yet this fell to 86.36% with the definitionally mismatching pairings (e.g. *host/mother*). Results of the matching condition are in line with those of Oakhill et al. (2005) who found accuracy of fillers to be uniformly high at around 95% across congruency conditions. However, this deterioration in accuracy on the mismatching pairs was driven by poorer accuracy on those involving definitionally male (77.81%) as opposed to definitionally female role names (94.91%). It appears that participants were interpreting certain male terms (e.g. *host, hero*) as

generically applicable to both sexes until they were alerted to the fact that they should be stricter in their linguistic definitions – this information was either signalled through the social feedback or by encountering the definitionally female counterpart to a male term that may previously have been presented e.g. once the term *hostess* has appeared, the term *host* is less likely to be interpreted generically. Therefore, as opposed to the gender stereotype bias evident in response to stereotype incongruent trials, poor performance on male definitionally mismatching trials suggests participants were responding in a more inclusive manner, accepting the male role terms as suitable referents to both sexes.

**Fillers - Response times.** The response time data tell a similar story. Average reaction times to definitionally matching word pairs were again faster ($M$ = 1048ms) than to definitionally mismatching word pairs ($M$ = 1094ms), with faster RTs to female word pairs over male word pairs in both the definitionally matching (999ms vs. 1048ms respectively) and mismatching cases (1044ms vs. 1145ms). This trend supports the accuracy data, with longer processing of male pairings likely to reflect participants' reflection over certain definitionally male terms which have female-specific counterparts and which should, therefore, be taken as male specific.

**Discussion**

Experiment 1 sought to investigate the influence of social consensus feedback on levels of gender stereotype application. Based on past research, it was hypothesised that a discrepancy between participants' responses and that of the perceived attitude of their peers would induce a feeling of discomfort, thus motivating participants to adapt their responding in line with their peer group (i.e. reduce stereotypic responding). While using social norm information as a strategy to reduce stereotype bias has proved successful in the past (Puhl et al., 2005; Stangor et al., 2001), it had remained untested in the field of gender stereotyping.

A significant improvement in accuracy to stereotype incongruent word pairs across blocks was found, thus identifying social consensus feedback as a useful stereotype-reduction strategy.

However, correct responses to incongruent word pairs remained significantly lower than to stereotype congruent and neutral pairs by the end of the current study, despite ample scope for further improvement. Similarly, while response times to all congruency conditions decreased significantly from Block 1 to Block 3 (and most dramatically in the case of stereotype incongruent pairings), participants remained slower to respond to stereotype incongruent pairings than stereotype congruent pairs by the end of the experiment. Interestingly, RTs to incongruent pairings did not initially improve when feedback was introduced in Block 2. Instead, it was when feedback was once again removed in Block 3 of the judgement trials that an acceleration of response times was evident. As the social feedback was conveyed in the form of a sentence, it is possible that participants took longer to process and digest the information, thus resulting in delayed changes to their patterns of responding.

Taken together, the accuracy and RT data suggest that presenting participants with social norm information is a useful means of attenuating the activation of spontaneous gender biases so as to result in lower levels of stereotype application. As a reminder, stereotype activation is an automatic process that results from increased cognitive accessibility of attributes associated with members of a particular social group, while stereotype application is typically under the conscious control of the perceiver and involves actual use of stereotypes in response to a group member (Kawakami et al., 2005). Therefore, participants succeeded in overcoming the spontaneous stereotype bias associated with the selected role nouns so as to ultimately judge that men can partake in traditionally female-biased roles and women can partake in male-biased roles. Through reference to the fact that gender stereotyping was not tolerated among their peer group, it appears that participants were motivated to adapt their responding and conform to the perceived behaviour of their peers. Given the successful use of social norm information as a means of stereotype reduction towards other minority groups in the past (e.g. racial minorities and those suffering from obesity), Experiment 1 provides further support in favour of this strategy, but now in the domain of gender stereotyping.

## Experiment 2

In Experiment 1, it was hypothesised that stereotype reduction was achieved through social compliance towards the perceived bias of a participant's peer-group. However, one further mechanism through which the social consensus feedback may have operated was by simply *alerting* participants to the issue of stereotype bias through the use of majority feedback (either in support of or in opposition to a participant's judgements). This majority feedback may have simply reminded participants that nowadays males can do jobs typically held by women and vice versa. Experiment 2 was therefore designed as a control experiment, aimed at differentiating between these two possibilities. In order to successfully distinguish between these two mechanisms, the design of Experiment 2 remained identical to that of Experiment 1 but with one modification – the feedback to critical, stereotype incongruent, word pairs was now centered on 50% (ranging from 35%-65%). This revised range of feedback was intended to suggest that people frequently *endorsed* stereotype biases, unlike Experiment 1 in which feedback implied that people rarely (2%-5% of the time) endorsed stereotypes. This form of feedback was now termed reverse social consensus feedback (RSCF). The issue under investigation was whether people (a) "comply" with this RSCF by becoming more like their allegedly stereotyped peer group, and thus fail to reduce levels of stereotypic responding across blocks or (b) whether feedback alerts participants to the issue of stereotype biases and leads them by a relatively indirect route to accept counter-stereotypes as possible, thus successfully reducing levels of stereotypic responding across blocks.

More specifically, if participants conform to the feedback provided, and maintain stereotype biases following the provision of RSCF in Block 2, no improvement in responding from Block 2 to Block 3 is anticipated. Conversely, if participants simply modify their behaviour once alerted to the issue of stereotype biases, an improvement in counter-stereotypic responding from Block 2 to Block 3 would be anticipated.

A pertinent issue for Experiment 2 was the range of feedback to be presented in response to

stereotype incongruent pairings. Essentially, a much wider range of feedback responses was now deemed necessary than the range of 4% used in Experiment 1 (where it was conveyed that 2%-5% of past participants endorsed stereotyping, and 95% to 98% of past participants rejected stereotyping). Although such a narrow feedback range was previously appropriate so as to strongly communicate that stereotyping was not supported, it was feared that this range would appear unrealistic if used in relation to stereotype endorsement i.e. it was not considered plausible to state that all previous participants rejected stereotype incongruent pairings within any given 4% range, especially around the mid-point of 50% e.g. 50%-54%, 51-55%, 49-53% etc. A greater feedback range of 35%-65% (i.e. 31%) was consequently selected for the current study[7].

**Method**

**Participants.** Thirty-three students (17 female, 16 male) from the University of Sussex took part in this experiment. Participants' ages ranged from 18 to 29 years (*M*: 19.27; *SD*: 2.54). They received either £6 or 4 course credits for taking part in the session which lasted approximately 45 minutes.

**Materials & Procedure.** Materials were identical to those of Experiment 4, but with the fictitious feedback updated so as to range from 35-65% for both correct (*yes*) and incorrect (*no*) responses to stereotype incongruent pairings. In this way, for half of the pairings the feedback quite strongly indicated that stereotype biases were being endorsed (e.g. stating that a figure between 50%-65% of people had responded that the terms *bricklayer* and *aunt* could *not* refer to one person), while for the other half the RSCF indicated that stereotype biases were being endorsed somewhat less often (for purposes of credibility, between 35% and 50% of the time). However, even this lower range of stereotype endorsement was much greater than the endorsement portrayed in Experiment 1 (2%-5%). As before, the exact figures that were used in conjunction with each of the stereotype incongruent pairings were randomly assigned within the specified range (of 35%-65%) but additionally distributed such that (a) no number appeared more than once and (b) the male

and female-biased incongruent pairings equally endorsed or rejected counter-stereotypic responding. Three distinct lists of these combinations were created, with the feedback values varied in each and used randomly across participants. The feedback provided in response to the other congruency conditions was consistent with that outlined in Experiment 1 (Appendix C).

Other details of the materials and procedure in Experiment 2 were identical to those of Experiment 1, aside from the debriefing. Participants were again informed of the aims of the experiment and reassured that the feedback provided was fictitious. However, it was further clarified that, although the feedback in this instance was expected to maintain the effects of stereotype biases, this experiment was in fact designed as a control condition for another study (Experiment 1) aimed at stereotype reduction, and that endorsement of gender stereotypes was not encouraged.

**Results**

Data trimming measures followed by two mixed-design ANOVAs were conducted as outlined in Experiment 1.

**Accuracy.** Evidence of stereotyping was first revealed with a main effect of Congruency, $F_1$ (1.04, 32.31) = 17.01, $p < .001$; $F_2$ (2, 32) = 60.68, $p < .001$. This effect was driven by significantly lower accuracy to stereotype incongruent word pairs ($M$ = 80.9%) than to stereotype congruent ($M$ = 98.25%) and neutral ($M$ = 96.7%) pairings. Importantly, there was no evidence of a significant Block by Congruency interaction, $F_1$ (4, 124) = 0.31, $p$ = .871; $F_2$ (4, 64) = 0.42, $p$ = .800. Furthermore, an inspection of Figure 3 suggests that there was no significant increase in accuracy of responses to stereotype incongruent word pairs from Block 1 to Block 3, $t_1$ (32) = 1.67, $p$ = .105; $t_2$ (23) = 1.39, $p$ = .178. Given this lack of improvement in accuracy scores across blocks, the data provide provisional evidence that participants complied with the perceived attitudes of their peer group, as opposed to attempting to overcome stereotypic responding upon being alerted to the issue of stereotype bias through the feedback provided.

**--insert Figure 3 about here--**

**Response times.** Response times for all errors of judgement were excluded from analysis (representing 10.49% of the data) along with extreme response times, below 150ms, and above 4,000ms (representing a further 2.88%), totalling a loss of 13.37% of the data. Again, the Participant by Block mean was calculated for each participant and data points 2.5 standard deviations above or below the Participant by Block mean were replaced with the relevant upper or lower cut off point (a further 4.21% of the data).

A main effect of Congruency was again found, $F_1$ (2, 62) = 21.84, $p < .001$; $F_2$ (2, 32) = 17.84, $p < .001$, with similarly fast RTs to stereotype congruent and neutral word pairs ($M$ = 760ms and 777ms respectively), followed by much slower judgements to stereotype incongruent pairings ($M$ = 876ms). There was again no interaction of Congruency by Block, $F_1$ (4, 124) = .72, $p$ = .578; $F_1$ (4, 64) = .96, $p$ = .434, with RTs decreasing significantly across blocks in each of the congruency conditions (see Figure 4): stereotype incongruent: $t_1$ (32) = 5.27, $p < .001$, $dz$ = .92; $t_2$ (23) = 7.72, $p < .001$, $dz$ = 1.58; stereotype congruent: $t_1$ (32) = 5.08, $p < .001$, $dz$ = .88; $t_2$ (23) = 7.60, $p < .001$, $dz$ = 1.55; neutral: $t_1$ (32) = 3.36, $p$ = .002, $dz$ = .59; $t_2$ (21) = 5.25, $p < .001$, $dz$ = 1.12.

**--insert Figure 4 about here--**

Finally, despite significantly slower RTs to the stereotype incongruent pairings in Block 1 compared to RTs of both congruent ($t_1$ (32) = 3.92, $p < .001$, $dz$ = .683; $t_2$ (23) = 8.76, $p < .001$, $dz$ = 1.79) and neutral word pairs ($t_1$ (32) = 3.82, $p$ = .001, $dz$ = .67; $t_2$ (44) = 6.73, $p < .001$, $d$ = 2.03.), only a significant difference between RTs of congruent and incongruent pairings remained by Block 3, $t_1$ (32) = 2.39, $p$ = .023, $dz$ = .42; $t_2$ (23) = 2.41, $p$ = .025, $dz$ = .49.

These improved RTs across blocks are relatively complicated to interpret in terms of the stated hypotheses. While the decrease may simply be due to practice effects, with participants consistently speeding up as the experiment progressed, they may alternatively provide evidence that participants are improving at the judgement task upon being alerted to stereotype biases with the

provision of feedback in Block 2. However, the latter of these two possibilities is deemed unlikely given a distinct lack of accompanying improvement in the accuracy data. Therefore, while the RT data do not directly provide support for the hypothesis that participants are conforming to perceived biases of their peers, when combined with the accuracy data it appears most likely that social compliance mechanisms are indeed the reason for which accuracy led to reduced stereotyping in Experiment 1.

**Fillers - Accuracy.** As in Experiment 1, average accuracy to the definitionally matching word pairs was higher than to the definitionally mismatching word pairs (95.10% vs. 84.60% respectively), with poorer accuracy in response to definitionally mismatching word pairs that contained a male-specific role name ($M$ = 75.96%) than a female-specific role name ($M$ = 93.23%). Once again it is hypothesised that this pattern of results is due to the generic interpretation of some of the definitionally masculine terms.

**Fillers - Response times.** RTs to definitionally matching word pairs were faster than to definitionally mismatching pairs (872ms vs. 944ms respectively). Also, faster RTs to female word pairs over male word pairs were found in both the definitionally matching (838ms vs. 906ms respectively) and mismatching cases (907ms vs. 980ms respectively). Again, these findings are in line with the accuracy data, as longer processing is likely to reflect participants' deliberation over terms which are masculine-specific by definition but frequently used in a more generic manner.

## Discussion

The aim of Experiment 2 was to establish the mechanism(s) through which social consensus feedback is likely to have succeeded as a stereotype reduction strategy in Experiment 1. While it was hypothesised that social compliance mechanisms were underlying the effects, an alternative possibility is that participants attempted to overcome stereotypic responding upon being alerted to the issue of stereotype biases through the variable feedback provided.

The results of Experiment 2 predominantly provide support for the first of these two

proposals, i.e. that stereotype change resulted from social compliance mechanisms. This conclusion was reached as accuracy of responses to stereotype incongruent pairings was not found to significantly improve across blocks, but remained in line with the perceived attitudes of participants' peers. Although RTs *were* found to significantly decrease across blocks (thus suggesting that participants were overcoming stereotypic responding upon being alerted to the issues of stereotype biases through the feedback provided), the lack of accompanying improvement in the accuracy data (and results of other work from the authors' lab, Finnegan et al., 2014.) suggests that this pattern of results simply stemmed from practice effects.

Although accuracy to stereotype incongruent word pairs was not found to significantly increase across blocks, neither did it decrease. This pattern of results was not surprising and echoes data reported in past research that attitudes are more difficult to influence in a negative direction than positive. For example, Puhl et al. (2005, Experiment 1) asked participants to estimate the percentage of obese people who possess 10 negative and 10 positive traits. These authors report a significant increase in positive trait ratings after participants received social feedback indicating past students had responded in this direction. However, trait ratings did not change in the unfavourable feedback condition in which participants learnt that other students attributed obese people with more negative trait ratings than positive. Similarly, Stangor et al. (2001) report that attitudes towards racial minorities were easier to influence in a positive direction than negative (although they concede this may have been due to issues of social desirability).

In conclusion, Experiment 2 suggests that social consensus feedback is likely to succeed as a stereotype reduction strategy through social compliance mechanisms as opposed to simple awareness of the issue of stereotype biases.

## General Discussion

The objectives of this research were to (a) investigate the efficacy of social consensus feedback as a strategy for overcoming gender stereotype application in a judgement task

(Experiment 1) and to (b) identify the mechanisms behind the success of this strategy (Experiment 2). While use of social norm information has previously proven successful as a strategy for overcoming prejudice in relation to racial minorities (Stangor et al., 2001) and those suffering from obesity (Puhl et al., 2005), it had remained untested in the field of gender stereotyping.

Experiment 1 found that both accuracy and response times of judgements to stereotype incongruent word pairs improved significantly following the provision of feedback in Block 2, and that this improvement was maintained in Block 3 when the feedback was again removed. However, despite this success, performance to incongruent pairings remained significantly poorer than to stereotype congruent and neutral pairings at the end of the experiment. As such, the provision of social consensus feedback did not succeed in wholly eradicating the stereotyping effect and adds to previous findings which show that stereotype biases are highly resistant to change (e.g. Oakhill et al., 2005; Reynolds et al., 2006).

Experiment 2 examined whether the social consensus feedback prompted compliance in participants towards the perceived attitudes of their peer-group, or simply alerted them to the issue of stereotype bias through the feedback provided, thereby inducing reduced levels of stereotyping. The RT data were ambiguous on this point as speeding up across blocks may have been due to practice effects as opposed to participants attempting to subvert stereotype biases. However, a corresponding lack of improvement in accuracy across blocks provides evidence for the former theory i.e. that it is indeed social compliance which drives behavioural change towards the perceived actions of one's peer-group upon the provision of social consensus information.

A point of concern with the reported experiments relates to Block 1 (pre-training) performance. Although both experiments were identical up to this point, performance was found to be somewhat variable at this early stage. For instance, Block 1 accuracy in Experiment 1 was 76.50% while Block 1 accuracy of Experiment 2 was 79.79% (a difference of 3.29%). As a consequence, although performance to incongruent pairings improved significantly across blocks in

Experiment 1 yet did not in Experiment 2, final Block 3 scores were very similar across studies (82.64% vs. 81.68% respectively). However, as we were primarily interested in participants' response to the consensus information (i.e. whether they would reduce or endorse stereotyping in line with the perceived attitude of their peers), we maintain that this issue is not crucial to the conclusions we have drawn.

But how do these findings sit within the existent stereotype reduction literature? The results of Experiment 1 fall in line with the claim of Prentice and Miller (1993) that participants will attempt to move their personal attitudes towards that of the perceived norm when they perceive their attitudes to be different from the normative attitude of their peer group. Furthermore, results provide support for Social Reality Theory (Hardin & Conley, 2000) and Group Norms Theory (Crandall et al., 2002; Kelman, 1958; Sherif & Sherif, 1953), which both posit that the human objective of social connection drives people to validate their experiences with others and to exhibit behaviours valued by admirable in-group members. Conversely, the findings do not support Deviance Regulation Theory (Blanton & Christie, 2003), which posits the rejection of perceived attitudes of peers as a means of self-definition.

Our findings also echo seminal work by Shiffrin and Schneider in the area of automatic and controlled processing. These authors argued that automatic responding can indeed be "unlearned" and that a newer response to a particular stimulus can come to dominate an old one (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). However, as automatic processes operate via a relatively enduring set of associative connections situated in long-term memory, the development of new automatic processes would require a considerable amount of consistent training to fully develop. In the field of stereotype reduction, this emphasis on training and repetition has also been advocated by Kawakami et al. (2000) whose stereotype negation training successfully led to reduced levels of automatic stereotyping towards skinheads and black students, with participants becoming increasingly efficient at overcoming stereotype activation across trials.

Kawakami and colleagues argue that cognitive changes may have resulted from the differential reinforcement and weakening of certain category-trait associations. The learning of new associations may have led to stereotype dilution, and in turn, reduced stereotype activation. This proposal seems equally plausible in terms of the social-feedback training, with participants creating stronger associations between previously weak category-member associations, and vice versa. They also suggest that motivational factors may have played a role in the success of their training i.e. through repeated activation of the goal 'to not stereotype', participants may have learned to spontaneously apply a self-regulatory process, a theory closely linked to the auto-motive model of Bargh and colleagues (Bargh, 1990; Bargh & Gollwitzer, 1994; Chartrand & Bargh, 1996). In this model, goals and motives must be represented in the mind in a way akin to that of other knowledge structures, and thus be capable of becoming automatically associated with representations that they are repeatedly paired with. Therefore, given the large number of 456 trials in our studies, it is likely that regulatory processes and the goal of stereotype-free responding became automated to a certain extent, thereby resulting in reduced levels of stereotype application.

The development of laboratory based interventions aimed at overcoming spontaneous stereotypes has important implications for the field of discourse processing. Much evidence currently supports the idea that gender is activated elaboratively once a stereotyped role noun is encountered (e.g. Oakhill et al., 2005; Pyykkönen, Hyönä, & van Gompel, 2010; Reynolds, et al., 2006) and that these gender biases affect language comprehension in a backward manner, while resolving anaphors that refer back to a stereotyped role name (e.g. builder…s/he) e.g. in self-paced reading tasks (e.g. Carreiras et al., 1996; Kennison & Trofe, 2003), eye-tracking (e.g. Duffy & Keir, 2004; Irmen, 2007; Kreiner et al., 2008) and electrophysiological studies (e.g. Osterhout, Bersick, & McLaughlin, 1997). However, if lab-based training can successfully help readers to overcome or control spontaneous stereotype biases in response to certain terms, and refrain from assigning gender in the absence of explicit definitional gender information, then the processing difficulties

that arise when the reader has to accommodate unexpected gender information into their mental model should be eliminated, and text comprehension improved.

Nevertheless, while previous research has documented the successful use of social feedback in overcoming prejudice and stereotyping, the success of this strategy proved somewhat limited in Experiment 1, as accuracy increased a relatively small amount across blocks (6.14%) and still lagged behind performance in other conditions. Reasons for this limited success relative to past research are unknown, yet may be attributable to plausibility of the feedback provided. While past studies often checked whether participants were suspicious about the feedback provided (e.g. Sechrist & Stangor, 2001; Stangor et al., 2001), we did not formally collect data on this issue (aside from during the pilot study of Experiment 2). As a result, it cannot be unequivocally ascertained that participants considered the feedback they received to truly reflect the attitudes of their peers. Indeed, past research suggests that participants who indicate suspicion about conformity are less, as opposed to more, likely to exhibit conformity effects (Stricker, Messick, & Jackson, 1967); a finding which may account for the relatively small magnitude of the social consensus training reported here.

It is also possible that differences in the current study design relative to past work may have contributed to differential effects of social consensus feedback across research. For example, dependent variables in previous work include trait ratings towards members of racial groups (Experiment 1, Sechrist & Stangor, 2001; Stangor et al., 2001), and those suffering from obesity (Puhl et al., 2005), intergroup helping behaviour towards African American versus White people (Sechrist and Milford, 2007), seating distance from an African American confederate (Experiment 1, Sechrist & Stangor, 2001), and a lexical decision task (Experiment 2, Sechrist & Stangor, 2001) as opposed to the judgement task used throughout this paper. Also, past work occasionally provided social consensus feedback and assessed dependent variables in the same experimental session (as in our research e.g. Sechrist & Stangor, 2001), while at other times a one week delay followed initial testing before the feedback was then provided and the dependent variable assessed on the same (e.g.

Experiments 1 and 3, Puhl et al., 2005; Experiments 1 and 3, Stangor et al., 2001) or different measures (e.g. Experiment 2, Puhl et al., 2005; Experiment 2, Stangor et al., 2001). Such differences in study design across existent research make it difficult to pinpoint exact reasons for past success in using peer-related interventions to influence stereotyping and prejudice, yet conversely illustrate the potential of such interventions to be effectively used across a broad range of methodologies and study designs. Future research should aim to identify conditions in which consensus information can be optimally used to tackle inequality across a variety of different social groups.

However, although evidence from Experiments 1 and 2 suggests that social consensus feedback is a useful means of stereotype reduction, in line with Stangor and colleagues (2001), it is not proposed that interventions should aim to indiscriminately modify stereotypes outside of the laboratory using false information about the opinions of others. That said, in cases where individuals incorrectly assume that stereotypic beliefs are widely shared or they over-estimate the negativity of stereotypes held by fellow group members (e.g. through the phenomenon of pluralistic ignorance, see Prentice and Miller, 1993), it is possible that providing people with accurate consensus information may be sufficient to generate stereotype change. As a result, it is important to further pinpoint the normative character of stereotype change and continue to build on the positive results already found with peer-related interventions.

References

Bargh, J. A. (1990). Goal and intent: Goal-directed thought and behavior are often unintentional. *Psychological Inquiry*, *1*(3), 248–251.

Bargh, J. A., & Gollwitzer, P. M. (1994). Environmental control of goal-directed actions and behavior. In W. Spaulding (Ed.), *Integrations of motivation and cognition: The Nebraska Symposium on Motivation* (Vol. 41, pp 71-124). Lincoln: University of Nebraska press.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*(6), 1142-1163.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828-841.

Blanton, H., & Christie, C. (2003). Deviance regulation: A theory of action and identity. *Review of General Psychology*, *7*(2), 115-149.

Carreiras, M., Garnham, A., Oakhill, J. V., & Cain, K. (1996). The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology, 49A*, 639–663.

Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, *71*(3), 464-478.

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: the struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359-378.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*(5), 800-814.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5-18.

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, *32*(4), 551–559.

Ford, T. E., & Ferguson, M. A. (2004). Social consequences of disparagement humor: A prejudiced norm theory. *Personality and Social Psychology Review*, *8*(1), 79–94.

Finnegan, E., Oakhill, J., & Garnham, A. (2014). Performance related feedback as a strategy to overcome gender stereotypes. Manuscript in preparation.

Gabriel, U., Gygax, P., Sarrasin, O., Garnham, A., & Oakhill, J. (2008). Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German. *Behavior Research Methods*, *40*(1), 206–212.

Garnham, A., Gabriel, U., Sarrasin, O., Gygax, P., & Oakhill, J. (2012). Gender representation in different languages and grammatical marking on pronouns: When beauticians, musicians, and mechanics remain men. *Discourse Processes*, *49*(6), 481–500.

Garnham, A., Oakhill, J., & Reynolds, D. (2002). Are inferences from stereotyped role names to characters' gender made elaboratively? *Memory & Cognition*, *30*(3), 439–446.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, *44*(2), 370–377.

Girden, E. R. (1992). *ANOVA: Repeated measures*. Newbury Park: Sage Publications.

Hamilton, S. (2008). *Automatic gender stereotyping, an ERP Investigation*. Unpublished masters thesis, University of Sussex, Brighton, UK.

Hardin, C. D., & Conley, T. D. (2000). A relational approach to cognition: Shared experience and relationship affirmation in social cognition. In G. Moskowitz (Ed.), *Cognitive social psychology: The Princeton symposium on the legacy and future of social cognition* (pp. 3-17), Hillsdale, NJ: Erlbaum.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 512–521.

Irmen, L. (2007). What's in a (role) name? Formal and conceptual aspects of comprehending personal nouns. *Journal of Psycholinguistic Research*, *36*(6), 431–456.

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and Social Psychology Bulletin*, *26*(8), 1002–1012.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*(5), 871-888.

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology*, *41*(1), 68–75.

Kelman, H. C. (1958). Compliance, identification, and internalization: Three processes of attitude change. *The Journal of Conflict Resolution*, *2*(1), 51–60.

Kennison, S. M., & Trofe, J. L. (2003). Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycholinguistic Research*, *32*(3), 355–378.

Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language*, *58*(2), 239–261.

LaFrance, M., & Woodzicka, J. A. (1998). No laughing matter:Women's verbal and nonverbal reactions to sexist humor. In J.Swim & C. Stangor (Eds.), *Targets of prejudice (pp. 61-80)*. San Diego: Academic Press.

Lassonde, K. A., & O'Brien, E. J. (2013). Occupational stereotypes: activation of male bias in a gender-neutral world. *Journal of Applied Social Psychology*. *43*(2), 387-396.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, *67*(5), 808-817.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology*, *65*(3), 469-485.

Monteith, M. J., Devine, P. G., & Zuwerink, J. R. (1993). Self-directed versus other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology*, *64*(2), 198-210.

Oakhill, J., Garnham, A., & Reynolds, D. (2005). Immediate activation of stereotypical gender information. *Memory & Cognition*, *33*(6), 972–983.

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*(3), 273–285.

Paluck, E. L. (2011). Peer pressure against prejudice: A high school field experiment examining social network change. *Journal of Experimental Social Psychology*, *47*(2), 350–358.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, *64*(2), 243-256.

Puhl, R. M., Schwartz, M. B., & Brownell, K. D. (2005). Impact of perceived consensus on stereotypes about obese people: a new approach for reducing bias. *Health Psychology*, *24*(5), 517-525.

Pyykkönen, P., Hyönä, J., & van Gompel, R. P. (2010). Activating gender stereotypes during online spoken language processing: evidence from Visual World Eye Tracking. *Experimental Psychology*, *57*(2), 126-133.

Reynolds, D. J., Garnham, A., & Oakhill, J. (2006). Evidence of immediate activation of gender information from a social role name. *The Quarterly Journal of Experimental Psychology*, *59*(05), 886–903.

Sanford, A. J. (1985). *Cognition and cognitive psychology.* London: Weidenfeld & Nicolson.

Schneider W., Eschman A., Zuccolotto A. (2002). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1-66.

Sechrist, G. B., & Milford, L. R. (2007). The influence of social consensus information on intergroup helping behavior. *Basic and Applied Social Psychology*, 29, 365–374.

Sechrist, G. B., & Stangor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. *Journal of Personality and Social Psychology*, *80*(4), 645-654.

Sherif, M., & Sherif, C. W. (1953). *Groups in harmony and tension; an integration of studies of intergroup relations*. New York: Harper.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127-190.

Stangor, C., Sechrist, G. B., & Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, *27*(4), 486–496.

Stricker, L. J., Messick, S., & Jackson, D. N. (1967). Suspicion of deception: Implications for conformity research. *Journal of Personality and Social Psychology*, *5*(4), 379-389.

Wittenbrink, B., & Henly, J. R. (1996). Creating social reality: Informational social influence and the content of stereotypic beliefs. *Personality and Social Psychology Bulletin*, *22*(6), 598–610.

Footnotes

[1]While this difference in bias ratings was not ideal, it was deemed more pertinent to choose the most strongly biased role nouns for each sex than to choose role nouns with matching degrees of typicality (as evidence of overcoming stereotyping to the strongest exemplars should logically extend to role nouns with a weaker bias rating).

[2]We use the term *critical* to refer to stereotype biased and neutrally rated items, and word pairs that include such an item.

[3]However, despite the fact that accuracy to male, definitionally mismatching word pairs was found to be relatively low in previous experiments by these authors (likely due to the generic interpretation of certain male-specific terms e.g. host, landlord), feedback continued to strongly suggest that such terms should be interpreted according to their definitional gender i.e. as being male-specific. For instance, if a mismatch pairing such as 'host/mother' was judged as acceptable, feedback stated that only 0-2% of people agreed with this response.

[4]No reliable main effects or interactions with this variable were found and so it is not considered further. Indeed, sex differences in performance were not anticipated in this study based on previous findings of Oakhill et al., 2005.

[5]A one-tailed *t*-test was used for this comparison (as it was anticipated that performance on the

incongruent pairings would improve after the social feedback training) while all remaining differences were examined using two-tailed tests. This procedure was also followed for the RT data.

[6]However, findings from another study from our lab (Finnegan et al., 2014) suggest that practice effects are also likely to have contributed to this decrease in RTs i.e. participants in a control condition who did not receive feedback after responding in Block 2 of the same judgement task were still found to naturally increase their speed of responding as the experiment progressed.

[7]A short pilot study was conducted to assess the credibility of the chosen feedback range. Eight students (all female) were administered just one block of judgement trials (used in Experiment 1), with feedback provided after each response. The newly constructed feedback centred on 50% for stereotype incongruent pairings, ranging from 35%-65% for both *yes* and *no* responses. On completion of the block of trials, participants were asked a number of questions about their experience of the task. Most importantly, on a scale of 1 (believable) to 5 (unbelievable) it was found that participants judged the RSCF to be "quite believable" ($M = 2$, $SD = 1.07$) and all participants reported feeling influenced by the feedback they received. Satisfied with these findings on the plausibility of the feedback provided, Experiment 2 was subsequently conducted using the same parameters. Note that the behavioural data from the pilot study was not analysed as we were simply interested in ascertaining whether participants found the fictitious feedback to be believable or not.

**APPENDIX A**

**Stereotyped role nouns used in Experiment 1 and 2**

| Male Stereotype | Female Stereotype | Neutral to Stereotype |
|---|---|---|
| Bricklayer | Beautician | Pedestrian |
| President | Fortune teller | Proof reader |
| Boxer | Au pair | Author |
| Mechanic | Secretary | Trainee |
| Football coach | Dressmaker | Neighbour |
| Lorry driver | Cleaner | Gynaecologist |
| Hunter | Flight attendant | Jogger |
| Factory manager | Social worker | Concert go-er |
| Electrician | Model | Relative |
| Pilot | Nurse | Office worker |
| Golfer | Chocolate lover | Artist |
| Politician | Birth attendant | Adolescent |

## APPENDIX B

### Filler (definitionally gendered) role nouns used in Experiment 1 and 2

Note some items were repeated across both conditions while others were used in one only.

| Male Definitional | | Female Definitional | |
|---|---|---|---|
| Policeman | Husband | Landlady | God mother |
| Groom | Landlord | Heroine | Policewoman |
| Postman | God father | Mistress | Grandmother |
| Salesman | Count | Spinster | Seamstress |
| Bachelor | Gigolo | Hostess | Geisha |
| Steward | Baron | Bride | Lesbian |
| Waiter | Fireman | Waitress | Matron |
| King | Grandfather | Princess | Baroness |
| Craftsman | Milkman | Mermaid | Nun |
| Prince | Host | Ballerina | Step mother |
| Son | Duke | Stewardess | Maid of honour |
| Sir | Best man | Milkmaid | Barmaid |

| | | | |
|---|---|---|---|
| Knight | Barman | Salesgirl | Wife |
| Master | Step brother | Duchess | Queen |
| Pope | Step father | Countess | Madam |
| Hero | Priest | Dame | Daughter |

**Appendix C**

**Fictitious feedback range: Social Consensus Feedback**

| | ***Yes* judgement** | ***No* judgement** |
|---|---|---|
| **Critical Items** | | |
| Neutral terms | 97-100% | 0-3% |
| Male/Female stereotype congruent terms | 97-100% | 0-3% |
| Male/Female stereotype incongruent terms | 95-98% | 5-2% |
| **Fillers** | | |
| Definitional gender match | 98-100% | 0-2% |
| Definitional gender mismatch | 0-2% | 98-100% |

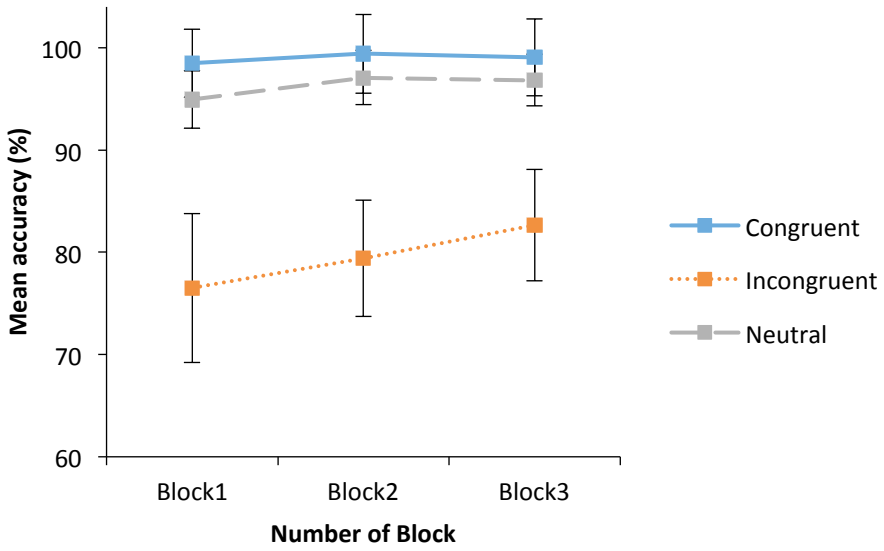


*Figure 1*. Experiment 1: Mean percentages of correct judgements to critical word pairs across blocks. The vertical axis begins at 60% while error bars indicate the 95% confidence intervals.

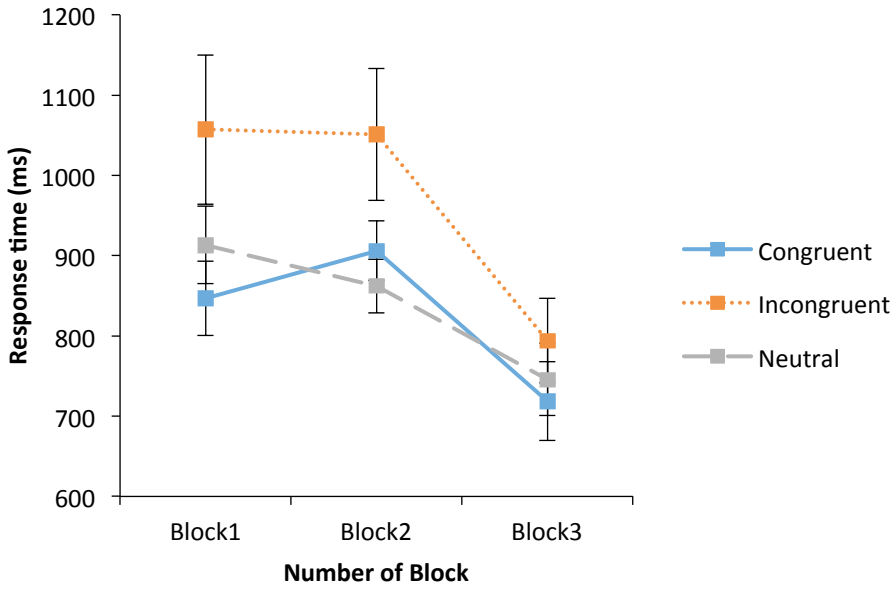


*Figure 2.* Experiment 1: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks. The vertical axis begins at 600ms while error bars indicate the 95% confidence intervals.

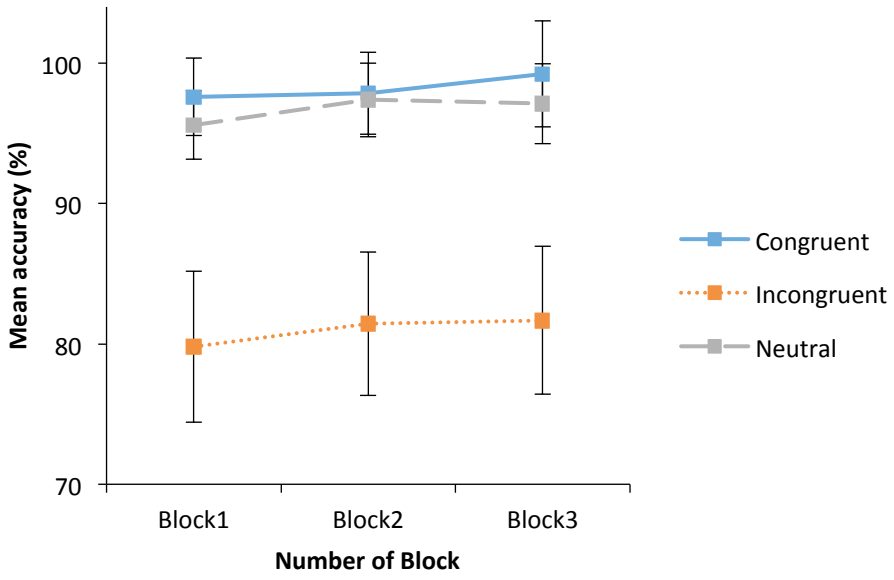


*Figure 3*. Experiment 2: Mean percentages of correct judgements to critical word pairs across blocks. The vertical axis begins at 70% while error bars indicate the 95% confidence intervals.

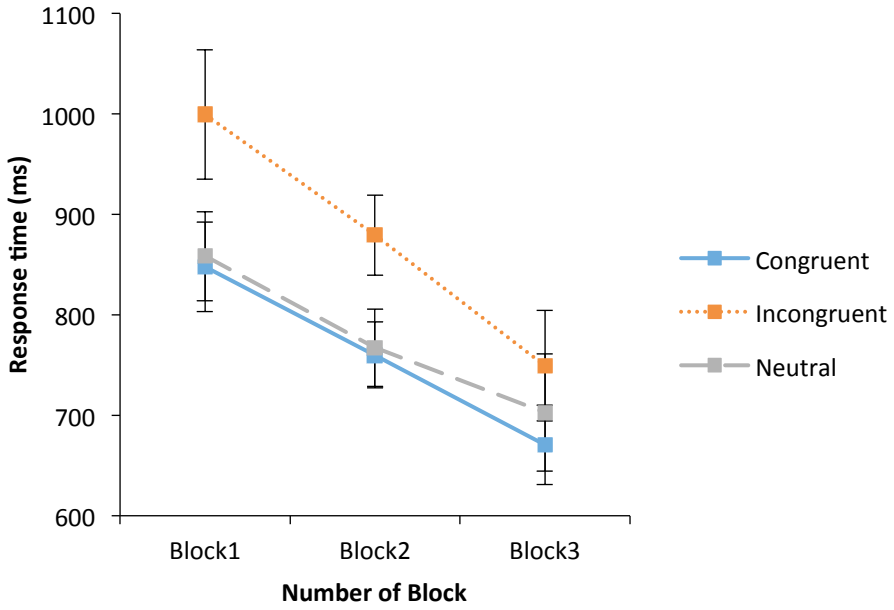


*Figure 4*. Experiment 2: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks. The vertical axis begins at 600ms while error bars indicate the 95% confidence intervals.